JUSTIN
Jurnal Sistem dan Teknologi Informasi
p-ISSN : 2460-3562 / e-ISSN : 2620-8989
DOI: 10.26418/justin.v8i2.36776
Vol. 8, No. 2, April 2020

Implementasi *Sentiment Analysis* Tanggapan Masyarakat Terhadap Pembangunan di Kota Pontianak

Lian Ardiani [a1], Herry Sujaini [a2], Tursina [a3]
[a]*Jurusan Informatika Universitas Tanjungpura*
Jl. Prof. Dr. H. Hadari Nawawi, Pontianak, Kalimantan Barat 78115
[1]lianardiani@gmail.com
[2]hs@untan.ac.id
[3]tursina@informatika.untan.ac.id

Abstrak

Sentiment analysis merupakan proses untuk memahami dan mengolah data tekstual secara otomatis untuk mendapatkan informasi sentimen yang terkandung dalam suatu kalimat opini. Pada umumnya masyarakat di zaman modern ini menuangkan dan mengekspresikan opininya ke media sosial terhadap berbagai topik, salah satu media sosial yang digunakan adalah *twitter*. Penelitian ini mencoba menganalisis *tweet* untuk dilakukan implementasi *sentiment analysis* terhadap opini masyarakat yang tertuang dalam *twitter*. Implementasi ini dilakukan dengan mengklasifikasikan *tweet* untuk mendapatkan informasi sentimen yang terkandung dalam tanggapan masyarakat, salah satu metode pengklasifikasian sentimen yaitu *naïve bayes*. Metode klasifikasi *naïve bayes* atau dikenal juga dengan teorema bayes memiliki ciri utama dalam asumsi opini yaitu menggunakan metode probabilitas dan statistik, teorema bayes menghitung nilai probabilitas tertinggi untuk klasifikasi sentimen. Jika suatu kata sering muncul dalam suatu dokumen maka diasumsikan bahwa kata tersebut merupakan kata penting dan diberikan nilai tertinggi, tapi jika kata muncul dalam berbagai dokumen maka kata tersebut bukanlah kata unik maka kata akan diberikan nilai rendah, dalam teorema bayes kata sendiri merupakan suatu unigrams dimana kata merupakan sentimen. Pengujian implementasi berbasis web menggunakan Bahasa Pemograman PHP menujukkan bahwa *tweet* dapat terklasifikasi secara otomatis. Data diklasifikasikan kedalam 3 (tiga) kelas yaitu, kelas positif, negatif dan netral. Proses dalam melakukan *sentiment analysis* dilakukan dengan melakukan *text processing*, setelah melewati tahapan *text processing* opini akan diklasifikasikan dengan metode klasifikasi *naïve bayes* ke dalam kelas positif, negatif atau netral. Secara keseluruhan hasil pengujian dengan menerapkan metode klasifikasi *naïve bayes* untuk mengolah informasi sentimen yang terdapat dalam suatu data *tweet* secara otomatis yang dikategorikan dalam tiga kelas yaitu positif, negatif dan netral dengan jumlah data latih sebesar 450 dan data uji sebesar 50, didapatkan akurasi sebesar 72%.

Kata kunci: *Sentiment analysis*, *naïve bayes*, teorema bayes, *twitter*, opini.

Implementation of Sentiment Analysis of Community Responses to Development in Pontianak City

Abstract

Sentiment analysis is a process to understand and process textual data automatically to get sentiment information contained in an opinion sentence. In general, people in modern times pour and express their opinions on social media on various topics; one of the social media used is Twitter. This study tries to analyze the tweet to be carried out the implementation of sentiment analysis of public opinion contained in Twitter. This implementation is carried out by classifying tweets to get sentiment information listed in public responses; one of the methods of classifying sentiments is naïve Bayes. The naïve Bayes classification method, also known as the Bayes theorem, has the main character in opinion assumptions, namely using the probability and statistical purposes, the Bayes theorem calculates the highest probability value for sentiment classification. If a word often appears in a document, then it is assumed that the word is an important word and is given the highest value, but if the name appears in various documents, then the word is not unique then the name will be given a low cost, in Bayes theorem the word itself is a unigram where words are sentiments. Testing the web-based implementation using the PHP Programming Language shows that tweets can be classified automatically. The data is classified into 3 (three) classes, i.e., positive, negative, and neutral classes. The process of doing sentiment analysis is done by doing text processing; after passing through the text, processing opinion will be classified with the naïve

Submitted 09-10-2019; Revised 27-03-2020; Accepted 08-04-2020

Bayes classification method into positive, negative, or neutral classes. Overall test results by applying the naïve Bayes classification method to process sentiment information contained in a tweet data automatically categorized into three categories, positive, negative and neutral with a total of 450 training data and test data of 50, obtained an accuracy of 72% .abstracts and keywords are made in two languages, Indonesian and English. Title letters use Times New Roman size 18. Abstract letters use Cambia Math, size ten and have a maximum of 300 words. The title and content of the abstract must provide the reader with information about the main content of the paper being written. The abstract content briefly describes the problem, objectives, methods/models used, results, and conclusions resulting from the research conducted. Abstracts may not contain pictures, equations, or library references. Avoid writing abbreviations without explanation. Keyword letters use Cambia Math size 10, which includes a maximum of six keywords. Keywords can be words or pairs of words that can describe the paper written.

Keywords: Sentiment analysis, naïve bayes, bayes theorem , twitter, opinion.

I. PENDAHULUAN

Informasi saat ini sangat mudah diperoleh dan dibagikan, karena begitu cepat perkembangan internet dengan adanya media sosial seperti salah satunya yaitu *twitter* yang memberikan kemudahan akses bagi pengguna internet untuk membagikan informasi dan membahas serta menuangkan opini terhadap berbagai topik. Opini yang tertulis pada media sosial mayoritas berupa data tekstual, opini tersebut dapat bersifat positif, negatif ataupun netral. Untuk mendapatkan informasi sentimen yang terkandung dalam opini tersebut maka dilakukanlah suatu analisis untuk mengevaluasi dan menilai data tekstual tersebut sehingga dapat menjadi sebuah informasi yang berharga, pengolahan analisis ini dikenal dalam dunia pemrosesan teks dengan *sentiment analysis*. *Sentiment analysis* merupakan proses untuk memahami dan mengolah data tekstual secara otomatis untuk mendapatkan informasi sentimen yang terkandung dalam suatu kalimat opini [1]. Penelitian di bidang *sentiment analysis* atau yang dikenal dengan *opinion mining* mulai marak dilakukan pada tahun 2002. *Sentiment analysis* acap kali digunakan untuk menilai suatu produk. *Opinion mining* bisa dianggap sebagai kombinasi antara *text mining* dan *natural language processing* [2].

Penelitian ini mencoba untuk mengklasifikasikan tanggapan atau sentimen masyarakat yang penerapannya terhadap pembangunan di Kota Pontianak. Banyak tantangan di dalam *sentiment analysis* salah satunya bagaimana menentukan suatu *tweet* bersentimen positif, negatif atau netral. Kedua adalah banyaknya pengguna internet dalam mengekspresikan opininya dengan gaya bahasa yang berbeda-beda dalam menulis *tweet*. Klasifikasi opini menjadi hal yang tidak mudah dimana juga pengguna *twitter* hanya dapat menulis postingan secara singkat karena jumlah penulisan karakter *twitter* dibatasi, sehingga terdapat banyaknya data *tweet* yang terdapat kalimat dengan kosakata dan tata bahasa yang tidak baku. Kemudian bagaimana suatu kalimat opini dianggap bersifat positif dan sebaliknya bahwa suatu opini dianggap bersifat negatif [3]. Untuk mengurangi *noise* yang terdapat pada *tweet* , dilakukan sebuah proses yaitu *text preprocessing*, Dalam *text processing* dilakukan beberapa tahapan yaitu berupa *case folding*, tokenisasi, *filtering*, dan *stemming*. Setelah dilakukan tahapan *text preprocessing*, data *tweet* akan diklasifikasikan kedalam kelas sentimen dengan menggunakan metode klasifikasi *naïve bayes*.

II. TINJAUAN PUSTAKA

A. Tanggapan dan Opini

Menurut Kamus Besar Bahasa Indonesia (KBBI), tanggapan merupakan apa yang diterima oleh pancaindra, pendapat, pandangan, sambutan dan reaksi. Menurut Kamus Besar Bahasa Indonesia (KBBI) opini adalah pendapat; pikiran atau pandangan, dan opini menurut *Webster's New Collegiate Dictionary* adalah suatu pandangan, keputusan atau taksiran yang terbentuk di dalam pikiran mengenai suatu persoalan tertentu.

Jadi dapat disimpulkan bahwa dalam memberikan suatu reaksi terhadap apa yang telah diamati dan dilihat serta dirasakan dengan pancaindra diberikan tanggapan berupa pendapat atau opini dalam mengekspresikan buah pikiran mengenai persoalan tertentu.

B. Sentiment Analysis

Sentiment analysis atau *opinion mining* mengacu pada bidang yang luas dari pengolahan bahasa alami, komputasi linguistik dan *text mining* yang bertujuan menganalisa pendapat, sentimen, evaluasi, sikap, penilaian dan emosi seseorang apakah pembicara atau penulis berkenan dengan suatu topik, produk, layanan, organisasi, individu, ataupun kegiatan tertentu. Tugas dasar dalam *sentiment analysis* adalah mengelompokkan polaritas dari teks yang ada dalam dokumen, kalimat, atau fitur/tingkat aspek apakah pendapat yang dikemukakan dalam dokumen, kalimat atau fitur entitas/aspek bersifat positif, negatif , atau netral [4]. Jadi, *sentiment analysis* berfokus pada pengolahan opini yang mengandung polaritas, yaitu memiliki nilai sentimen positif, negatif atau netral.

Berdasarkan sumber datanya *sentiment analysis* terbagi menjadi 2 kelompok yaitu, *coarse-grained sentiment analysis*, *sentiment analysis* yang dilakukan pada level dokumen, secara garis besar fokus utama dari *sentiment analysis* jenis ini adalah menganggap seluruh isi dokumen sebagai sebuah sentimen positif dan sentimen negatif. *Fined-grained sentiment analysis* adalah *sentiment analysis* yang dilakukan pada level kalimat, fokus utamanya adalah menentukan sentimen pada setiap kalimat. *Sentiment analysis* dapat diklasifikasikan kedalam kelas sentimen bersifat positif, negatif dan netral [5].

1) Sentimen Positif: Menurut Kamus Besar Bahasa Indonesia (KBBI) sentimen positif merupakan reaksi atau sikap yang meningkatkan nilai seseorang atau sesuatu.

Dari struktur kalimat dalam penelitian ini kalimat bersentimen positif sebagai contoh "Senang sekali dengan adanya beberapa taman dan ruang baca, semoga bisa membuat kota Pontianak lebih bersinar".

2) Sentimen Negatif: Menurut Kamus Besar Bahasa Indonesia (KBBI) sentimen negatif merupakan reaksi atau sikap yang menurunkan nilai seseorang atau sesuatu, jadi kalimat bersentimen negatif akan menyebabkan penyurutan nilai pandang terhadap sesuatu, sehingga membentuk tren down. Umumnya kalimat bersentimen negatif ditandai dengan penggunaan kata negasi. Negasi merupakan sesuatu yang dikenal dalam semua bahasa dan biasanya negasi digunakan untuk mengubah polaritas dari suatu pernyataan [6]. Dari struktur kalimat bersentimen negatif sebagai contoh "Rencana pemerintah merenovasi Taman Budaya Kalimantan Barat tak pernah terealisasi. Tak pernah di perbaiki sejak di bangun tahun 1982 cc @BangMidji agar di perhatikan".

3) Sentimen Netral: Kata netral sendiri dalam Kamus Besar Bahasa Indonesia (KBBI) berarti tidak berpihak, kalimat bersentimen netral merupakan ekspresi kalimat yang tidak bersifat positif maupun negatif. Dari struktur kalimat bersentimen netral adalah sebagai berikut "Mengunjungi karnaval khatulistiwa di Pontianak #KalbarMenangTotal".

C. Text Preprocessing

Text preprocessing merupakan tahap awal dimana data teks dengan *noise* akan dikurangi agar dapat diolah lebih lanjut, untuk mengurangi *noise* tersebut. Tahap *text preprocessing* ini mencakup semua rutinitas, dan proses untuk mempersiapkan data yang akan digunakan pada operasi *knowledge discovery* sistem *text mining*. Data tekstual akan diproses dalam beberapa tahapan *text preprocessing* yaitu *case folding*, tokenisasi, *filtering* dan *stemming*. Dengan dilakukannya *text preprocessing* akan terbentuk *dataset* bersih, *dataset* yang terbentuk dari proses ini akan memudahkan dalam pemrosesan sistem.

4) Case Folding: Case folding dilakukan untuk mengubah setiap karakter didalam teks menjadi huruf kecil. Tidak semua kata dalam teks konsisten dalam penggunaan huruf kapital disinilah tujuan dilakukan *case folding* untuk mengkonversi setiap karakter dalam kata menjadi huruf kecil [6].

5) Tokenisasi: Tokenisasi merupakan proses pemecahan kata pada suatu teks ke dalam satuan kata. Tokenisasi dilakukan untuk menghasilkan kumpulan kata yang berdiri sendiri, tokenisasi memecah teks yang semula berupa kalimat menjadi kata-kata. tokenisasi menghilangkan delimeter seperti titik (.), koma (,), spasi, dan karakter angka yang ada pada kata tersebut [7].

Dalam penelitian ini tokenisasi dilakukan untuk memecah kata, serta melakukan penghapusan delimeter beserta karakter angka bersama *tweet entity* seperti *hashtag*, *retweet* dan *mention*.

6) Filtering: Filtering merupakan proses dalam *text preprocessing* setelah tokenisasi, *filtering* dilakukan untuk untuk mengambil kata penting hasil tokenisasi. Pada tahap filtering kata akan ditentukan apakah akan digunakan atau dibuang. Proses dalam *filtering* dalam membuang kata-kata yang tidak digunakan atau *stopword* terdapat dalam *bag of words stoplist* [8].

Stopword merupakan daftar kata-kata yang tidak mempresentasikan isi dari suatu dokumen teks, *stopword* dilakukan untuk meghilangkan kata atau *term* yang tidak memiliki arti. Daftar *stoplist* akan dibuat sebelum melakukan proses *stopword removal*, jika kata-kata terdapat dalam daftar *stoplist*, maka kata tersebut akan dihapus, sehingga kata-kata yang tersisa akan dianggap kata yang mencirikan isi suatu dokumen.

7) Stemming: Stemming merupakan proses mengubah kata menjadi bentuk dasarnya. *Stemming* dilakukan untuk meyeragamkan bentuk kata. Tujuan dari proses *stemming* adalah menghilangkan imbuhan-imbuhan baik itu berupa prefiks, sufiks, maupun konfiks yang ada pada setiap kata [9]. *Stemming* dalam penelitian ini dilakukan berdasarkan aturan morfologi bahasa Indonesia.

D. Klasifikasi

Klasifikasi merupakan proses menentukan atau menilai suatu objek untuk memasukkannya ke kelas-kelas yang telah ditentukan. Klasifikasi teks dapat didefinisikan sebagai proses untuk menentukan suatu dokumen teks ke dalam suatu kelas tertentu [10]. Dalam penelitian ini klasifikasi dilakukan untuk memprediksi orientasi *tweet* kedalam kelas yang sudah ditentukan yaitu positif, negatif atau netral. Proses klasifikasi dalam *sentiment analysis* dapat dilakukan dengan berbagai algoritma, salah satunya yaitu *naïve bayes*.

Naïve bayes merupakan salah satu metode yang digunakan pada *text mining* yang sederhana tetapi memiliki keakuratan yang tinggi dalam mengklasifikasi [11]. *Naïve bayes* memiliki waktu klasifikasi yang singkat sehingga mempercepat proses *sentiment analysis*. Metode klasifikasi *naïve bayes* merupakan metode klasifikasi berdasarkan probabilitas dan teorema *bayes* (aturan *bayes*) dengan asumsi indenpendensi yang kuat.

Dalam algoritma *naïve bayes* setiap dokumen dipresentasikan dengan pasangan atribut "$x_1, x_2, x_3, ..., x_n$" dimana x_1 adalah kata pertama, x_2 adalah kata kedua dan seterusnya. Probabilitas dari semua dokumen (*posterior probability*) suatu dokumen pada suatu kategori dapat dihitung dengan menggunakan persamaan berikut [12],

$$V_{MAP} = \underset{V_j eV}{\arg\max} \frac{P(X_1, X_2, ..., X_n | V_j) P(V_j)}{P(X_1, X_2, ..., X_n)} \pi \quad (1)$$

Klasifikasi teks menggunakan *naïve bayes* dilakukan dengan memaksimalkan nilai dari persamaan, asumsi kemandirian bersyarat yang "naif" memegang peranan. Untuk penyebut $P(X_1, X_2, ..., X_n | V_j)$ nilainya konstan atau sama untuk semua kategori (V_j). Sehingga persamaan dapat ditulis sebagai berikut,

$$V_{MAP} = \underset{V_j eV}{\arg\max} P(X_1, X_2, ..., X_n | V_j) P(V_j) \quad (2)$$

Persamaan diatas dapat disederhanakan menjadi,

$$V_{MAP} = \underset{V_j eV}{\arg\max} \prod_{i=1}^{n} P(X_i \mid V_j) P(V_j) \quad (3)$$

Keterangan:

V_j = Kategori tweet j = 1, 2, 3,…n.

Dimana dalam penelitian ini j_1 = kategori *tweet* sentimen positif, j_2 = kategori *tweet* sentimen negatif, dan j_3 = kategori *tweet* sentimen netral.

$P(X_i \mid V_j)$ = Probabilitas x_i pada kategori V_j.

$P(V_j)$ = Probabilitas dari V_j.

Untuk nilai dan dihitung pada saat pelatihan dimana persamaannya adalah sebagai berikut:

$$P(V_j) = \frac{|\ docs\ \ j\ |}{|\ contoh\ |} \quad (4)$$

$$P(X_i \mid V_j) = \frac{n_k + 1}{n + |\ kosakata\ |} \quad (5)$$

Keterangan:

$|\ docs\ j\ |$ = Jumlah dokumen setiap kategori j.

$|\ contoh\ |$ = Jumlah dokumen dari semua kategori.

n_k = Jumlah frekuensi kemunculan kata, untuk menghindari nilai 0 maka pembilang ditambahkan 1.

n = Jumlah frekuensi kemunculan kata dari setiap kategori.

$|\ kosakata\ |$ = Jumlah semua kata dari semua kategori.

Ringkasan dari algoritma dengan metode klasifikasi *naïve bayes* adalah sebagai berikut [13]:

1. Proses pelatihan. Input adalah data yang sudah diketahui kategorinya
   a. $|\ kosakata\ |$ ← himpunan semua kata yang unik dari dokumen-dokumen
   b. Untuk setiap kategori V_j lakukan :
      - $|\ docs\ j\ |$ ← himpunan dokumen-dokumen yang berada pada kategori V_j.
      - Hitung $P(V_j)$.
      - Untuk setiap x_i pada $|\ kosakata\ |$ lakukan:
        1. Hitung $P(X_i \mid V_j)$.
2. Proses Klasifikasi. Input adalah dokumen yang belum diketahui kategorinya
   a. Hasilkan V_{MAP} dengan menggunakan $P(X_i \mid V_j)$ dan $P(V_j)$ yang telah diperoleh dari pelatihan.

E. Dataset

Dataset yang akan digunakan berasal dari media sosial *twitter* berupa *tweet* sebagai data tekstual. *Dataset* merupakan kumpulan data latih dan data uji, dari data ini akan dihitung akurasi berdasarkan persentase jumlah data latih dan data uji. Tahap pertama adalah melakukan pelatihan terhadap data opini yang telah diketahui kategorinya, dan tahap kedua adalah proses klasifikasi dokumen yang belum diketahui kategorinya. *Dataset* yang digunakan berfokus pada *tweet* berbahasa Indonesia yang membahas tentang pembangunan di Kota Pontianak.

F. Recall dan Precission

Recall dan *precision* digunakan untuk mengukur efektivitas sistem temu kembali informasi. Nilai dari *recall* dan precission biasanya ditunjukan dalam persen (%).

$$\text{Recall} = \frac{\text{jumlah dokumen relevan yang terpanggil (a)}}{\text{jumlah dokumen relevan yang ada di dalam database (a + c)}} \text{x}100 \quad (6)$$

$$\text{Precision} = \frac{\text{jumlah dokumen relevan yang terpanggil (a)}}{\text{jumlah dokumen yang terpanggil dalam pencarian (a + b)}} \text{x}100 \quad (7)$$

TABEL I
TABEL RELEVANSI

	Relevant	**Not Relevant**	**Total**
Retrieved	a (hits)	b (noise)	a+b
Not Retrieved	c (misses)	d (reject)	c+d
Total	a+c	b+d	a+b+c+d

Keterangan:

a (*hits*) : Dokumen yang relevan.

b (*noise*) : Dokumen yang tidak relevan.

c (*misses*) : Dokumen relevan yang tidak ditemukan.

d (*reject*) : Dokumen tidak relevan yang tidak ditemukan.

III. METODOLOGI PENELITIAN

Sentiment analysis merupakan salah satu cabang penelitian *text mining*. Secara khusus tujuan *text mining* dapat dibagi menjadi dua yaitu pengkategorisasian data teks (*text categorization*) dan pengelompokkan data teks (*text clustering*). Dalam pengkategorisasian, *text mining* dipergunakan sebagai alat untuk menemukan kategori yang sesuai dengan kelas yang ditentukan (*supervised learning*), sedangkan pengelompokkan dalam *text mining* berfungsi sebagai alat untuk mengelompokkan data teks berdasarkan kesamaan karakteristik, dan *clustering* dapat digunakan untuk memberikan label pada kelas yang belum diketahui (*unsupervised learning*) [14]. Tujuan utama pada penelitian ini adalah pengkategorisasian teks (*text categorization*), karena penelitian ini melakukan pengkategorian sebuah *tweet* apakah bersentimen positif, negatif atau netral.

C. Pengumpulan Data

Data yang digunakan dalam penelitian ini berupa tanggapan masyarakat mengenai pembangunan Kota Pontianak. Data dibagi menjadi data latih dan data uji. Proses pembuatan *dataset* didapatkan dari data yang telah diambil dari media sosial *twitter* berupa *tweets*, data berupa data tekstual. Data yang diterima kemudian akan dibuat data latih dan data ujinya, dimana data latih merupakan data yang telah diketahui kategorinya, sedangkan data uji merupakan data yang belum diketahui kategorinya. Data latih sendiri merupakan kumpulan data yang telah dibagi sesuai dengan kategorinya secara manual, masing-masing data latih akan dibagi ke dalam kelas sentimen positif, negatif dan netral.

D. Perancangan dan Pembuatan Sentiment Analysis

Pada proses ini akan dilakukan dalam 2 tahapan:

8) Text preprocessing: merupakan tahap dimana teks akan diseragamkan bentuk katanya agar dapat dipersiapkan menjadi data yang akan diolah selanjutnya. Tahapan *text preprocessing* meliputi *case folding*, tokenisasi, *filtering* dan *stemming*. Gambar.1 menunjukkan tahapan pada *text preprocessing*.

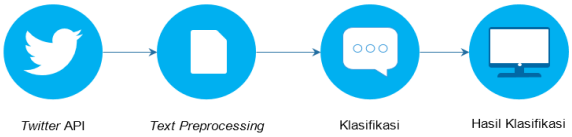
Gambar.1 Tahapan *text preprocessing*

Case Folding adalah proses penyeragaman dengan mengubah semua huruf dalam dokumen menjadi huruf kecil. Tokenisasi dalam penelitian ini merupakan tahapan dalam memecah *string* atau input terhadap suatu teks berdasarkan tiap kata yang menyusunnya dan menghilangkan URL, *@mention* dan *hashtag* yang ada pada dokumen (*tweets*) beserta menghilangkan delimeter-delimeter seperti tanda titik (.), koma (,), spasi dan karakter angka dan tanda baca yang ada pada dokumen (*tweets*). *Filtering* yaitu tahap mengambil kata-kata penting dari hasil tokenisasi atau merupakan tahap menghilangkan kata-kata yang tidak mengandung makna. *Filtering* dapat dilakukan dengan menghilangkan *stopwords*. *Stemming* merupakan pemetaan kata menjadi bentuk kata dasarnya.

9) Proses klasifikasi: dalam penelitian ini menggunakan metode klasifikasi *naïve bayes*.

E. Arsitektur Sistem

Sistem yang dibangun adalah sistem yang dapat digunakan dalam melakukan *sentiment analysis* pada tanggapan masyarakat terhadap pembangunan di Kota Pontianak pada twitter. Sistem bekerja dengan melakukan *crawling tweet* dengan memanfaatkan *twitter* API, yang nantinya akan diproses oleh sistem untuk mendapatkan nilai sentimen yang terdapat dalam *tweet* dengan melewati tahapan *text preprocessing*, kemudian setelah itu akan dilakukan klasifikasi sentimen untuk mendapatkan hasil klasifikasinya. Adapun diagram arsitektur sistem pada Gambar.2.



Gambar. 2 Arsitektur sistem

IV. HASIL PENELITIAN

A. Pengumpulan Dataset

Data berupa *tweet* yang berhubungan dengan kota Pontianak yang terkait dengan penelitian yang dilakukan. Proses pengumpulan data yang dilakukan dengan melakukan *crawling* manual pada *twitter search*, dari hasil yang didapatkan *dataset* yang digunakan berjumlah 500 *tweets*. Setelah *dataset* dibangun, masing-masing akan dibagi menjadi dua yaitu data uji dan data latih, dimana data latih dilakukan pengelompokkan menjadi 3 (tiga) kelas untuk dilabel sesuai jenis sentimen mana yang merupakan *tweet* bersentimen positif, negatif atau netral. Masing-masing berjumlah 150 data per kategori kelas data latih dan 50 data digunakan sebagai data uji dapat dilihat pada Tabel 2.

TABEL II
RINCIAN DATASET

Jenis Sentimen *Tweet*	**Jumlah**
Latih	450
Uji	50
Total *Dataset*	500

Contoh dataset yang dikumpulkan pada Gambar.3.

> Hari ini kekantor @BPJSTKinfo Pelayanan cepat petugasnya ramah dan ruangan dingin, meskipun cuaca pontianak panas. #Pontianak @BPJSTK_Mobile

Gambar. 3 Contoh *tweet*

B. Labelling

Setelah dataset telah terpenuhi, langkah selanjutnya melakukan pelabelan pada setiap *tweets* tersebut. Proses pelabelan dilakukan dengan memilih orientasi data dari *tweet* apakah positif, negatif dan netral. Pelabelan data *tweet* dilakukan secara manual. Pelabelan dilakukan untuk membentuk data latih, dan dilakukan juga pelabelan pada data uji untuk menghitung akurasi dari hasil sentimen yang didapat secara otomatis.

C. Text Preprocessing

1) Case Folding: Proses *case folding* dilakukan untuk mengubah semua huruf dalam dokumen menjadi huruf kecil. Gambar.4 merupakan proses *case folding* yang dilakukan.

> hari ini kekantor @bpjstkinfo pelayanan cepat petugasnya ramah dan ruangan dingin, meskipun cuaca pontianak panas.

Gambar. 4 Hasil tweet melewati tahapan *case folding*

2) Tokenisasi: Dari hasil tokenisasi kalimat yang sudah dipecah menjadi kata akan dimasukkan ke dalam array. Tabel 3 merupakan proses hasil tokenisasi yang telah dilakukan.

TABEL III
HASIL *TWEET* MELEWATI TAHAPAN TOKENISASI

hari	ini	kekantor	pelayanan	cepat
petugasnya	ramah	dan	ruangan	dingin
meskipun	cuaca	pontianak	panas	

3) Filtering: Filtering dilakukan untuk membersihkan kata-kata hasil tokenisasi dengan penggunaan *stopwords removal*. Pada penelitian ini *stoplist* disesuaikan dengan *tweet text*. Proses *filtering* yang dilakukan ditunjukkan pada Tabel 4.

TABEL IV
HASIL *TWEET* MELEWATI TAHAPAN *FILTERING*

hari	kekantor	pelayanan	cepat	petugasnya
ramah	ruangan	dingin	cuaca	pontianak
panas				

4) Stemming: Tahapan dari proses *stemming* yang dilakukan dengan menggunakan algoritma Arifin dan Setiono. Tahapan dari proses *stemming* dengan algoritma Arifin dan Setiono adalah sebagai berikut [15]:

a. Pemeriksaan semua kemungkinan bentuk kata. Setiap kata diasumsikan memiliki 2 awalan (prefiks) dan 3 akhiran (sufiks).
Prefiks 1 + Prefiks 2 + Kata dasar +Sufiks 3 + Sufiks 2 + Sufiks 1
Jika dalam kata yang diperiksa tidak memiliki imbuhan sebanyak imbuhan seperti formula di atas, maka imbuhan yang kosong atau tidak ada tersebut diberi tanda x untuk prefiks dan diberi tanda xx untuk sufiks.

b. Pemotongan dalam Algoritma ini dilakukan secara berurutan sebagai berikut:
AW : AW (Awalan)
AK : AK (Akhiran)
KD : KD (Kata Dasar)
- AW I, hasilnya disimpan pada pe1 (prefiks 1)
- AW II, hasilnya disimpan pada pe2 (prefiks2)
- AK I, hasilnya disimpan pada su1 (sufiks 1)
- AK II, hasilnya disimpan pada su2 (sufiks 2)
- AK III, hasilnya disimpan pada su3 (sufiks 3)

Pada setiap tahap pemotongan di atas diikuti dengan pemeriksaan di kamus apakah hasil pemotongan itu sudah berada dalam bentuk dasar. Kalau pemeriksaan ini berhasil maka proses dinyatakan selesai dan tidak perlu melanjutkan proses pemotongan imbuhan lainnya.

c. Akan tetapi, apabila sampai pada pemotongan AK III, belum ditemukan dalam kamus, maka akan dilakukan proses kombinasi. Kata dasar yang dihasilkan dikombinasikan dengan imbuhan-imbuhannya dalam 12 konfigurasi berikut:
- KD
- KD + AK III
- KD + AK III + AK II
- KD + AK III + AK II + AK I
- AW I + AW II + KD
- AW I + AW II + KD + AK III
- AW I + AW II + KD + AK III + AK II
- AW I + AW II + KD + AK III + AK II + AK I
- AW II + KD
- AW II + KD + AK III
- AW II + KD + AK III + AK II
- AW II + KD + AK III + AK II + AK I

Sebenarnya kombinasi a, b, c, d, h, dan l sudah diperiksa pada tahap sebelumnya, karena kombinasi ini adalah hasil pemotongan bertahap tersebut. Dengan demikian, kombinasi yang masih perlu dilakukan tinggal 6 yakni pada kombinasi-kombinasi yang belum dilakukan (e, f, g, i, j, dan k). Tentunya bila hasil pemeriksaan suatu kombinasi adalah 'ada', maka pemeriksaan pada kombinasi lainnya sudah tidak diperlukan lagi. Pemeriksaan 12 kombinasi ini diperlukan, karena adanya fenomena *overstemming* pada algoritma pemotongan imbuhan. Kelemahan ini berakibat pada pemotongan bagian kata yang sebenarnya adalah milik kata dasar itu sendiri yang kebetulan mirip dengan salah satu jenis imbuhan yang ada. Dengan 12 kombinasi itu, pemotongan yang sudah terlanjur tersebut dapat dikembalikan sesuai posisinya.

Setelah tahapan *filtering* akan dilakukan *stemming*, tahapan akhir dari *text preprocessing* dan setelah *stemming* data *tweet* akan diolah ke tahapan selanjutnya. *Stemming* yang dilakukan ditunjukkan pada Tabel 5.

TABEL V
HASIL *TWEET* MELEWATI TAHAPAN *STEMMING*

hari	kantor	layan	cepat	petugas
ramah	ruang	dingin	cuaca	pontianak
panas				

D. Pengujian

Pengujian yang dilakukan dengan membuat sistem berbasis *web* yang memanfaatkan *twitter* API yang berfungsi untuk mengambil data *tweet* secara terkini untuk dilakukan pengklasifikasian secara langsung ke dalam kategori kelas sentimen yang telah ditentukan yaitu positif, negatif dan netral. Proses klasifikasi menggunakan metode *naïve bayes*. Pemanfaatan *twitter* API dengan menggunakan sampel pencarian "@bangmidji" dengan hasil yang didapat pada Gambar 5.

Gambar. 5 Hasil pengujian *tweet* dengan menggunakan *twitter* API

Twitter API membatasi pengambilan data *tweet* sesuai dengan ketentuan pada halaman tentang *standard* API *twitter* yang bisa dilihat pada sumber dari situs *twitter* (https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets.html). Dimana ketentuan pada pengambilan *tweets* dengan memanfaatkan *standard twitter* API dibatasi hanya 15 *tweet*s selama 7 hari kebelakang

Pada penelitian ini juga dilakukan pengujian terhadap klasifikasi sentimen yang telah dilakukan. Perbandingan dataset yang dilakukan yaitu data latih 450 dan data uji 50 dengan metode klasikasi *naïve bayes*. dengan skenario yang dapat dilihat pada Tabel 6.

TABEL VI
TABEL SKENARIO *DATASET*

Data	Positif	Negatif	Netral
Latih	150	150	150
Uji	17	17	16
TOTAL	500		

Tweet pada data uji akan diprediksi oleh aplikasi pengujian yang telah dibuat, untuk melakukan pengecekan akurasi. Data uji sebelumnya telah telah dikategorikan secara manual ke dalam kelas-kelas yang telah ditentukan lalu akan dicek akurasinya. Tabel 7 menunjukkan data uji yang akan diklasifikasi untuk dihitung akurasinya.

TABEL VII
DATA UJI

DATA UJI		Prediksi			Total
		Negatif	Positif	Netral	
Label Manual	Negatif	12	1	4	17
	Positif	1	13	3	17
	Netral	3	2	11	16
TOTAL		16	16	18	50

Hasil Perhitungan *recall* yang didapat dari pengujian terhadap data uji yang telah dilakukan masing-masing sebesar negatif 70,59%, positif 76,47% dan Netral 68,75 dengan total *recall* keseluruhan dari data uji sebanyak 50 dan hasil prediksi benar sebanyak 36 didapatkan *recall* sebesar 72% dapat dilihat pada Tabel 8.

TABEL VIII
HASIL PERHITUNGAN *RECALL* DATA UJI

Kelas	Data Benar	Kelas Data Uji	Recall
Negatif	12	17	70,59%
Positif	13	17	76,47%
Netral	11	16	68,75%

Perhitungan *precission* dari data uji yang telah terprediksi oleh sistem yang telah dibuat pada Tabel 9, dimana hasil didapatkan masing-masing negatif 75%, positif 81,25% dan netral 61,12% dengan total keseluruhan *precission* dengan data uji sebanyak 50 dan data yang terprediksi benar sebanyak 36 didapatkan nilai *precission* sebesar 72%.

TABEL IX
HASIL PERHITUNGAN *PRECISSION* DATA UJI

Kelas	Data Benar	Jumlah Prediksi	*Precission*
Negatif	12	16	75%
Positif	13	16	81,25%
Netral	11	18	61,12%
Total	36	50	72%

Sekenario juga dilakukan untuk melihat pengaruh *dataset* terhadap hasil akurasi yang diperoleh pada pengujian ini dengan menguji data uji dengan beberapa kali proses pelatihan dimana data uji masing masing perkategori berturut-turut dengan jumlah data latih yang sama yaitu sebanyak 50 dan skenario pengujian pertama dengan data latih masing-masing kelas kategori sebanyak 150 didapatkan akurasi sebesar 72%, skenario kedua dengan data latih sebanyak 100 masing-masing kelas kategori didapatkan akurasi sebesar 66% dan skenario ketiga dengan data latih sebanyak 50 masing-masing kelas kategori didapatkan akurasi sebesar 52%. Hasil akurasi dari skenario pengujian yang telah dilakukan dapat dilihat pada Tabel 10.

TABEL X
HASIL PENGUJIAN

Data Latih			Data Uji	Akurasi
Positif	Negatif	Netral		
150	150	150	50	72%
100	100	100	50	66%
50	50	50	50	52%

E. Analisis Hasil Pengujian

Analisis hasil pengujian dilakukan untuk mengetahui apakah implementasi yang dibuat dapat mengklasifikasi *tweet* tentang pendapat masyarakat terhadap pembangunan Kota Pontianak secara otomatis. Analisis pengujian dilakukan untuk guna mengevaluasi kinerja sistem dalam klasifikasi.

Dari hasil pengujian didapatkan bahwa pengujian yang dilakukan dapat melakukan klasifikasi *tweet* secara otomatis dengan memanfaatkan *twitter* API untuk mendapatkan data *tweet* agar dapat dilakukan proses klasifikasi secara otomatis.

Perbandingan *dataset* yang dilakukan yaitu data latih sebanyak 450 dan data uji sebanyak 50 dengan metode klasikasi *naïve bayes* dan kategori kelas klasifikasi ke dalam tiga (3) kelas yaitu positif, negatif dan netral. Hasil pengujian pengklasifikasian dengan metode *naïve bayes* didapatkan akurasi sebesar 72 % , *precisson* 72% dan *recall* sebesar 72%. Diketahui juga dari pengujian yang telah dilakukan bahwa jumlah data latih mempengaruhi akurasi dalam pengklasifikasian.

V. KESIMPULAN

Adapun kesimpulan dari penelitian yang telah dilakukan adalah bahwa Pengujian yang dilakukan dapat mengklasifikasikan data tweet ke dalam sentimen positif, negatif atau netral secara otomatis. Jumlah data latih

mempengaruhi performasi klasifikasi sentimen beserta akurasinya. Metode klasifikasi naïve bayes dengan sekenario dataset, 450 data latih dan 50 data uji menghasilkan akurasi sebesar 72%, recall 72% dan precission 72 % .

Beberapa hal yang sarankan yaitu penambahan data latih dalam meningkatkan ketepatan pengkategorian data tweet, dan untuk menangani klasifikasi dapat menggunakan metode lain selain metode klasifikasi naïve bayes seperti metode SVM.

DAFTAR PUSTAKA

[1] A. V. Sudiantoro and E. Zuliarso, "Analisis Sentimen Twitter Menggunakan Text Mining Dengan Algoritma Naive Bayes Classifier," *Pros. SINTAK 2018*, 2018.

[2] [2] I. F. Rozi, S. H. Pramono, and E. A. Dahlan, "Implementasi Opinion Mining (Analisis Sentimen) untuk Ekstraksi Data Opini Publik pada Perguruan Tinggi," *Electr. Power, Electron. Commun. Control. Informatics Semin.*, 2012.

[3] [3] J. Jotheeswaran and Y. S. Kumaraswamy, "Opinion Mining Using Decision Tree Based Feature Selection Through Manhattan Hierarchical Cluster Measure," *J. Theor. Appl. Inf. Technol.*, 2013.

[4] [4] B. U. Manalu, "Analisis Sentimen pada Twitter Menggunakan Text Mining Skripsi Boy Utomo Manalu," *Teknol. Inf. Fak. Ilk. UNSUT*, 2014.

[5] [5] Falahah and D. Dwiki Adriadi Nur, "Pengembangan Aplikasi Sentiment Analysis Menggunakan Metode Naïve Bayes (Studi Kasus Sentiment Analysis dari media Twitter)," *Semin. Nas. Sist. Inf. Indones.*, 2015.

[6] [6] S. Adi, "Perancangan Klasifikasi Tweet Berdasarkan Sentimen dan Fitur Calon Gubernur DKI Jakarta 2017," *J. Inform. Pelita Nusant.*, vol. 3, no. 1, 2018.

[7] [7] D. Pakpahan and H. Widyastuti, "Aplikasi Opinion Mining dengan Algoritma Naïve Bayes untuk Menilai Berita Online," *J. Integr.*, 2014.

[8] [8] E. E. Pratama and B. R. Trilaksono, "Klasifikasi Topik Keluhan Pelanggan Berdasarkan Tweet dengan Menggunakan Penggabungan Feature Hasil Ekstraksi pada Metode Support Vector Machine (SVM)," *J. Edukasi dan Penelit. Inform.*, 2015.

[9] [9] F. Z. Tala, "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia," *M.Sc. Thesis, Append. D*, 2003.

[10] [10] A. F. Hidayatullah and M. R. Ma'arif, "Penerapan Text Mining dalam Klasifikasi Judul Skripsi," *Semin. Nas. Apl. Teknol. Inf. Agustus*, 2016.

[11] [11] A. Pramono, R. Indriati, and A. Nugroho, "Sentiment Analysis Tokoh Politik Pada Twitter," *Semin. Nas. Inov. Teknol. UN PGRI Kediri, 22 Februari 2017*, 2017.

[12] [12] B. Kurniawan, S. Effendi, and O. S. Sitompul, "Klasifikasi Konten Berita dengan Metode Text Mining," *J. Dunia Teknol. Inf.*, 2012.

[13] [13] N. W. S. Saraswati, "Text Mining dengan Metode Naïve Bayes Classifier dan Support Vector Machines untuk Sentiment Analysis," 2011.

[14] [14] A. Aprilianti P, "Sentiment Analysis dengan Naive Bayes untuk Melihat Persepsi Masyarakat Terhadap Batik pada Jejaring Sosial Twitter," in *Seminar Nasional Matematika dan Pendidikan Matematika UMS Tahun 2015*, 2015.

[15] [15] D. Novitasari, "Perbandingan Algoritma Stemming Porter dengan Arifin Setiono untuk Menentukan Tingkat Ketepatan Kata Dasar," *STRING (Satuan Tulisan Ris. dan Inov. Teknol.*, 2017.